How to make R, PostGIS and QGis cooperate for statistical modelling duties: a case study on hedonic regressions

Olivier Bonin

How to make R, PostGIS and QGis cooperate for statistical modelling duties: a case study on hedonic regressions

Olivier Bonin, Université Paris Est – IFSTTAR – LVMT

Hedonic regressions used to model housing prices are a good example of statistical modelling that is demanding towards software: housing databases are generally large (typically several hundred thousands of records), and many spatial characteristics must be computed for each housing location. An hedonic model for housing prices [2, 3] is generally written as:

$$p_i = \Sigma \alpha_j x_{ij} + \Sigma \beta_j y_{ij} + \Sigma \gamma_j z_{ij} + \varepsilon_i$$

where x_{ij} denotes structural characteristics of housing i (e.g. size, number of rooms, presence of an elevator, etc.), y_{ij} the neighbourhood characteristics (e.g. presence of rapid transits in the vicinity), z_{ij} the market position characteristics (e.g. distance to employment, distance to the city centre), p_i the housing prices (or more generally Box-Cox transforms of these prices), and ε_i a Gaussian error term.

The structural characteristics are found in housing price databases. The location of each housing consists in a postal addresses or in geographical coordinates. The spatial characteristics y_{ij} and z_{ij} are computed with the help of spatial queries as well as with network analyses. Thus it is necessary to be able to load road networks and public transportation networks and perform classical network analyses such as shortest path computation. It is also necessary to manage several layers of spatial information such as the location of employment centres and of amenities.

To select pieces of software to perform our modelling duties, we have to keep in mind that we want to navigate within two distinct worlds: the world of statistics and the world of geomatics. On one hand, a statistical software is mandatory to perform model estimation. R [4] is probably the best candidate for this task, even compared to commercial systems, given that our modelling might include spatial regression techniques or multi-level modelling [1]. On the other hand, a geographical information system (or a spatially-aware database system) is useful to perform spatial analysis (though R can perform part of these tasks thanks to some of its libraries). Here, the number of mature open source geographical information systems is relatively high, and several projects have reached the state to be able to cope with the hundred thousands of points associated to the location of dwellings. And globally, as a large amount of data is involved, it may prove useful to store and access it through a database management system. PostGIS[1] is a very good and easy to use spatially-aware RDBMS. Moreover, the spatial index in PostGIS will be helpful to perform spatial queries on hundred thousands of points. While R and PostGIS are obvious choices, the choice of the candidate geographical information system is very open. QGis[2] has been selected in the present study because it has the ability to connect to both PostGIS and R natively or with the help of readily available plugins, and has decent mapping capabilities (Figure 1).

---

[1] http://postgis.refractions.net/
[2] http://www.qgis.org/

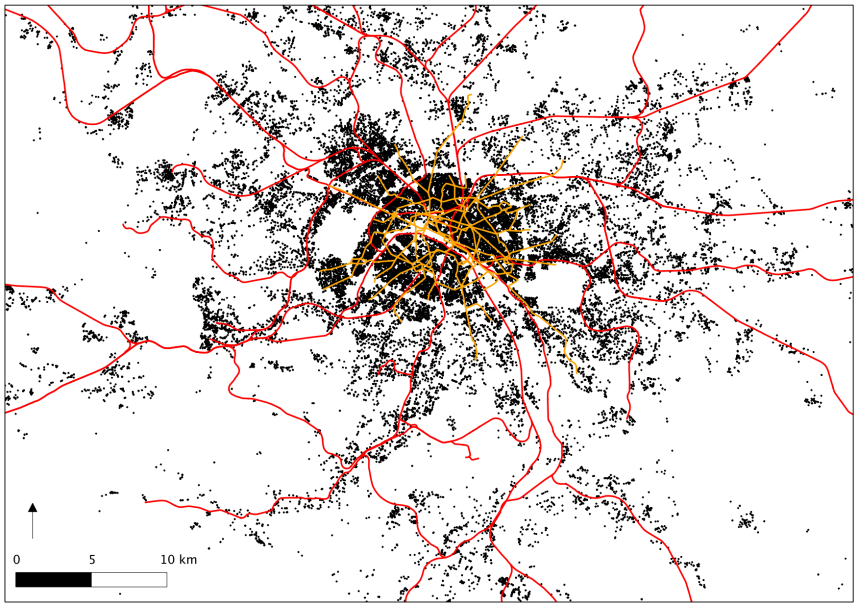

Figure1: Rapid transit network and housing locations in the Ile-de-France region (map drawn with QGis, data from the BIEN database)

We wanted our system to be able to run on Windows, GNU/Linux and Mac OS X, which is the case for the three softwares. However, in our setup with a PostGIS server on GNU/Linux and R and QGis on Mac OS X, we have discovered that some of the required libraries required to make software communicate could be a bit tricky to install and configure on Mac OS X.

We describe in this communication how we managed to make R, PostGIS and QGis work together to assist us in our econometric modelling of housing prices. We have evaluated different combinations of these three pieces of software that we have used extensively during the course of this research, to finally converge to the system that we are currently using. We were concerned mainly with performance (as large amounts of data are involved), spatial query capabilities, easiness to compute new spatial indicators and to take them into account in the statistical modelling, and finally spatial visualisation of statistical results and map production. As it is often the case with open source software, several possibilities exist to connect the pieces of software two by two. We tried some of them and give an insight on the results of our experiments.

We chose PostGIS as a central repository for all our tabular and spatial data, and R as statistical system. To our knowledge, R can be connected to PostGIS either by "RODBC"[3] or by "Rdbi" and "RdbiPgSQL". Rdbi (hosted on BioConductor[4] and thus accessible through this repository) proved to be much faster than RODBC in our case, so that we have settled on it. Moreover, having ODBC work on Mac OS X requires to compile the PostGIS driver and to configure the connexion by hand. To connect QGis to PostGIS was performed with QGis native driver. We have also connected QGis directly to R for testing purposes. This connexion can be established by the "manageR" plugin (and probably other plugins). It did work, but did not proved to be very useful: because of the size of the housing price database, loading data from files in QGis exhibited horrible

[3] http://cran.r-project.org/web/packages/RODBC/index.html
[4] http://bioconductor.org/

performance compared to PostGIS. The performance of the final system is quite good. Data can be exchanged between R and PostGIS in both directions with decent times, and from PostGIS to QGis for visualisation purposes. We almost never had to exchange data from Qgis to PostGIS, as all spatial queries were directly performed in PostGIS by SQL queries.

Our return on experience on these three pieces of open source software deals with statistical analysis, and more precisely the modelling of the influence of accessibility on housing prices. These kinds of models can require somewhat sophisticated statistical methods when spatial auto-correlation of residuals must be controlled and taken care of. Of utmost importance to us was the possibility to visualise our results on maps, so that we focus here on the visualisation duties.

The typical modelling process is to import data from PostGIS into R, estimate models, add the model error term into the data frame used for modelling and then update the PostGIS database with the new values. It is clear from Figure 1 that mapping the error term at the point level leads to hard to read maps, because of point density and because several dwellings share the same coordinates as the result of the (documented) geocoding process of dwellings in the BIEN database. Thus we choose to work on aggregate error terms, at the IRIS level or at the town level. Figure 2 is an example of such a map at the IRIS level on Ile-de-France.

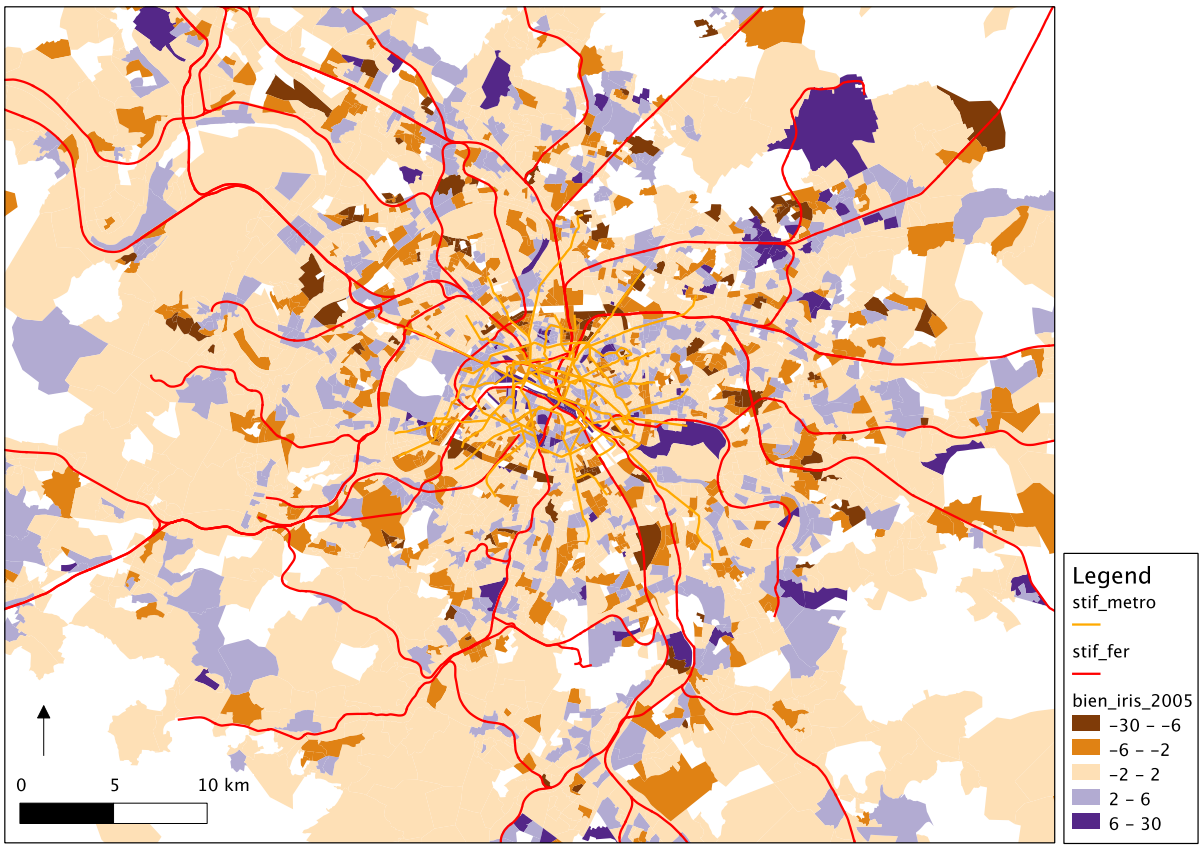

Figure2: residuals of one of our hedonic models on housings aggregated at the IRIS level in Ile-de-France (map done with QGis, data from the BIEN database)

However, with our progressive discovering of R spatial handling facilities, we use more and more R and PostGIS without QGis. Indeed, we are able to load shapefiles, perform analyses and thematic maps with the help of the “sp” and the “maptools” libraries. The shapefiles describing public transit networks, employment centres or administrative boundaries are small enough to be efficiently handled by R. Thus, we obtain quicker maps of error terms, probably cruder than the ones we can elaborate with QGis, but sufficient for analysis and even for publication.

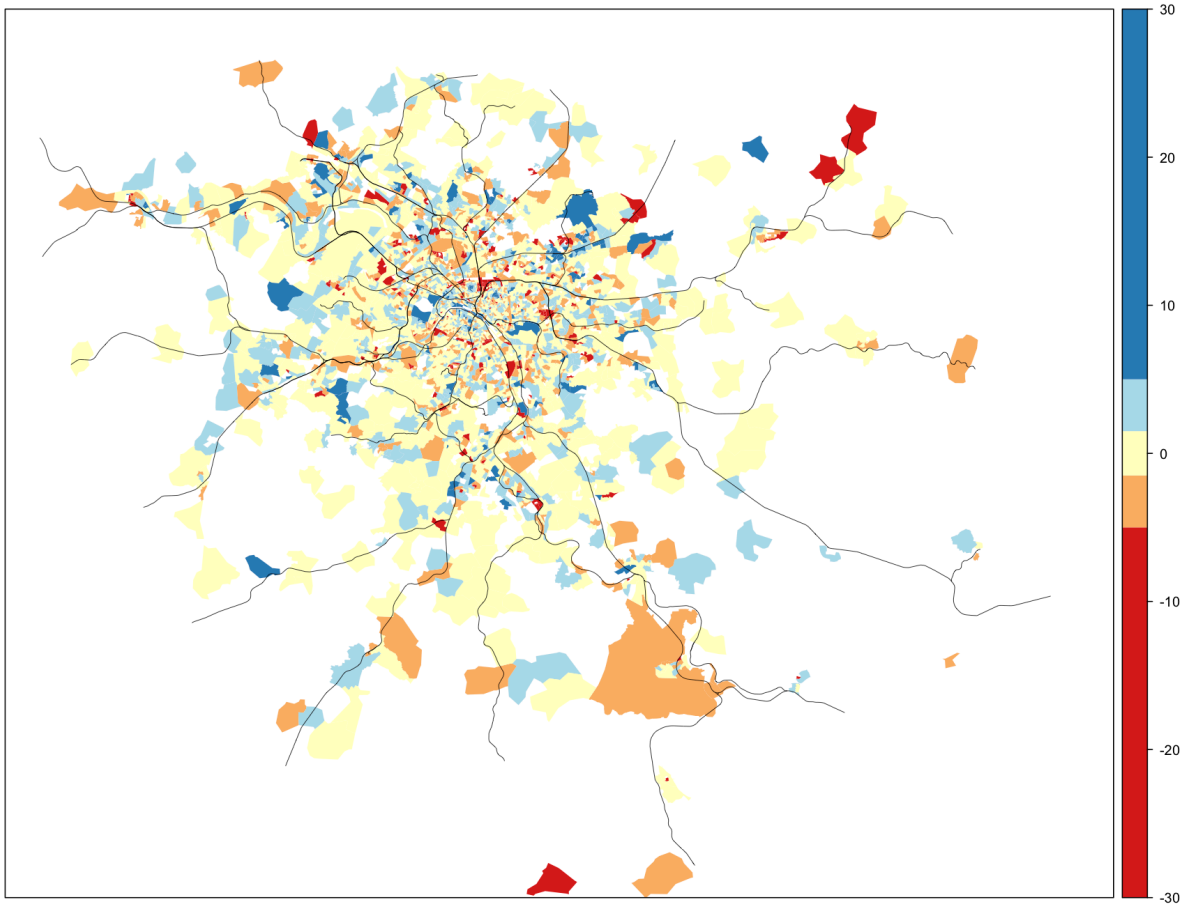

Figure 3: residuals of one of our hedonic models on housings aggregated at the IRIS level (map done in R with the spplot() method, data from the BIEN database)

The most striking conclusion of our experiment is that, despite the strong requirements of our research in spatial analysis, the geographical information system appeared to be the most dispensable piece of software. QGis showed irritating limitations when working on PostGIS tables (by comparison to working on shapefiles), and its visualisation module worked decently but showed real ergonomics problem when tuning cartographic representation. Of course there might have been better choices than QGis, but our experiments with other geographical information systems were not better.

We were able to progressively put aside the geographical information system in our setup because of the excellent spatial processing capabilities of PostGIS, and to the development of a large number of spatial functions inside R, including the ability to load shapefiles and compute spatial queries. Our conclusion on working on the interface

between statistics and geomatics is that statistical systems such as R are bridging the gap between both worlds. In our opinion, geographical information systems have to work a great deal towards statistical analysis to regain the interest of researchers performing quantitative analyses in social sciences.

References

[1] BONIN, O. Urban location and multi-level hedonic models for the lle-de-France region, *ISA International Conference on Housing*, Glasgow, 2009.

[2] KAIN, J. F., AND QUIGLEY, J. M. Measuring the value of house quality. *Journal of the American Statistical Association* 65(330): 532-548, 1970

[3] ROSEN, S. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* 82(1): 34-55, 1974.

[4] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2009.